GATE: improving the computational efficiency.

S. Staelens, [a,*] J. De Beenhouwer, [a] D. Kruecker, [b] L. Maigne, [c] F. Rannou, [d] L. Ferrer, [e] Y. D'Asseler, [a] I. Buvat, [f] I. Lemahieu, [a]

[a] *UGent-ELIS; St-Pietersnieuwstraat,41; B-9000 Gent*
[b] *Institute of Medicine-Forschungszemtrum Juelich; D-52425 Juelich; Germany*
[c] *Département de Curiethérapie-Radiothérapie; Centre Jean Perrin; F-63000 Clermont-Ferrand; France*
[d] *Departamento de Ingenieria Informatica; Universidad de Santiago de Chile; Santiago; Chile*
[e] *INSERM U601; CHU Nantes; F-44093 Nantes; France*
[f] *INSERM U678 UPMC; CHU Pitié-Salpêtrière; F-75634 Paris; France*

Abstract

GATE is a software dedicated to Monte Carlo simulations in Single Photon Emission Computed Tomography (SPECT) and Positron Emission Tomography (PET). An important disadvantage of those simulations is the fundamental burden of computation time. This manuscript describes three different techniques in order to improve the efficiency of those simulations. Firstly, the implementation of variance reduction techniques (VRTs), more specifically the incorporation of geometrical importance sampling, is discussed. A relative figure of merit was calculated for a standard setup, showing an efficiency enhancement of $5 - 15$ by using this technique. After this, the newly designed cluster version of the GATE software is described. The experiments have shown that GATE simulations scale very well on a cluster of homogeneous computers for the case of low sensitivity (SPECT) setups and that an optimum can be derived for high sensitivity (PET) experiments using this cluster package of the GATE software. Finally, an elaboration on the deployment of GATE on the EGEE (Enabling Grids for E-Science in Europe) Grid will conclude the description of efficiency enhancement efforts. The latter has shown to be efficient but depending on the queuing policy of the site accepting the jobs. The three aforementioned methods improve the efficiency of GATE to a large extent and make realistic patient-specific overnight Monte Carlo simulations achievable.

Key words: Monte Carlo, GATE, variance reduction, cluster, Grid

1. Introduction

A new software, GATE [1], was designed as an upper layer for the Geant4 nuclear physics code and was tuned for use in nuclear medicine, more specifically to fulfill its role as a simulation platform for PET and SPECT incorporating all Geant4 features. An important disadvantage of GATE Monte Carlo simulations is the fundamental burden of computation time. This manuscript will describe three different techniques in order to improve the efficiency of those simulations. Firstly we will discuss the implementation of VRTs, more specifically the incorporation of geometrical im-

* Corresponding author: Dr. S. G. Staelens

portance sampling. Secondly, the newly designed cluster version of the GATE software will be described. Finally, an elaboration on the deployement of GATE on the EGEE Grid will conclude the description of the efficiency enhancement efforts.

2. Methods

Benchmark simulations were performed to estimate the simulation time for realistic PET and SPECT nuclear medicine setups. These benchmarks have been described in full detail in [1]. The computing time for the PET benchmark averaged around twelve hours on a 1.0 GHz processor. This corresponds to 852 generated and tracked events per second and 16 simulated coincidence detections per second. Calculation time for the SPECT case was eleven hours on a 1 GHz processor, resulting in 417 generated and tracked events per second and 0.83 detections per second or 1.2 second per detection. The reasons for the high computation time in the SPECT simulations with a collimator are twofold. The total number of registered counts in SPECT is less than 0.02 % of the generated events because the collimator in front of the crystal stops most of the incoming photons and registers all processes present. This effect is most pronounced in high resolution collimator setups. Secondly, GATE tracks every photon through every object of the experiment, and a typical Low Energy High Resolution (LEHR) collimator for instance consists of 161120 individual air holes.

2.1. *Variance reduction techniques: methodology*

Geometrical importance sampling is a VRT based on the crude criterion that only photons with a high detection chance should be tracked. Photons are increasingly split into exact copies with lowered weights as the distance to a detector decreases. Photon paths leading away from a detector are less likely to result in detection and therefore these photons are subject to Russian roulette in order to increase the simulation efficiency. Particles following a path leading away from a detector undergo a statistical test which results in a higher weight if passed [2]. Geometrical importance sampling combined with Russian roulette introduces branches into the particle history. Simply adding all hits in a detector crystal would lead to a completely wrong simulated histogram. Therefore a new track history has been developed within Gate [3]. It keeps a log of all tracks and their weights generated by GEANT4 and tracks are added where necessary to accomodate splitting and Russian roulette. The efficiency gain of importance sampling is inversely related to the sensitivity of the detector and despite the complex detangling and increased tracking overhead, it can result in a 5 to 15-fold increase over analog simulations. The incorporation of importance sampling has been validated extensively. An efficiency comparison of this newly incorporated VRT with analog GATE simulations has been performed on a ^{67}Ga energy spectrum. Activity and acquisition time were kept constant while simulation time and number of detections are the parameters of interest. The efficiency will be defined as the number of detections/second times the quality factor (QF) [4]. The QF indicates the variation of the weights of the detected particles. A large spread in the weights could result in a high tally bin due to few large weight counts, but with a low QF. It is calculated by the following equation:

$$QF = (\sum weight)^2/(detections * \sum weight^2) \quad (1)$$

We hereby assume that the previous equation holds true for setups with few detected particles per simulated events as in [4]. The current implementation of importance sampling copies the time stamp which causes it currently to be limited to SPECT only while respecting the virtual clock philosophy for time management. In PET more complex sorting algorithms for coincidences, scatter fractions and random rates are needed which are mostly based on unique time window information. Moreover, in PET, gamma pairs are created which complicates the current track detangling implementation.

2.2. *Running GATE on a cluster: methodology*

In Monte Carlo simulations the amount of interprocess communication is small, and usually only at process start-up and termination. A distributed computing approach for running the highly scalable GATE experiments in a cluster of computers is an appropriate solution to reduce the overall computing time. The approach that will discussed in this manuscript is platform independent in the sense that the simulations are partitioned in virtual time slices so that the user obtains a number of fully resolved independent job execution macros accompanied by a platform specific submit file. The implementation we describe is fully automatic and requires no interaction from the user [5].

The parallelized simulations are made up of 3 steps : job splitting, the actual simulations (on a number of CPUs) and file merging. The most natural, simple and general scheme for splitting PET and SPECT simulations is the time-domain decomposition approach, in which the length of the experiment is split into a number of equally long smaller experiments. This approach does not involve any approximation nor simplification. Measures like random rates, scatter fractions, and system deadtime will be effectively the same as in a single-node run. The input to the job splitter are the GATE scripts, parameters and command-line options. A Random Number Generator (RNG) provides statistically independent seeds for the output which is a collection of non-parameterized (fully resolved) macro files. The job splitter also provides a submit file for supported cluster platforms such as Condor and openMosix, to facilitate the startup of the simulation. Finally, a split file is generated that contains all information about the partitioned simulation to facilitate the merging of the output files. Since GATE does not allow any volume movement during data acquisition, virtual time slices are used for the time-domain decomposition. The output merger takes then finally as input the ROOT output files from the parallelized simulations and uses the split file to merge them.

To test the efficiency enhancement, the GATE benchmarks were run. The cluster used was based on openMosix and consisted of 38 nodes with 17 dual XEON 2.4Ghz processors and 21 dual XEON 2.8Ghz processors, each with 2GB of memory. The benchmark series were executed using the cluster with an increasing number of CPUs.

Efficiency was hereby defined as an acceleration factor based on Amdahl's law, AF, being:

$$AF = \frac{T_s + T_p}{T_s + T_p/p} \qquad AE = \frac{T_s + T_p}{T_m + T_s + T_p/p} \quad (2)$$

where p is the number of CPUs , and T_s and T_p are the serial and parallel application times, respectively [6]. T_s was measured as the time necessary to process the GATE scripts that set up the simulations. T_p was calculated from the total simulation time and T_s. A more approporiate estimate (AE) of the acceleration factor can be made by inclusion of the merging time T_m into Amdahl's law, as in (2).

2.3. *Grid: Methodology*

EGEE [7] is part of the Grid initiatives launched by the European Union to structure the Grid infrastructures in Europe and build upon the already established GEANT infrastructure. GATE is a pilot biomedical application in this project. GATE simulations benefit from the geographically distributed grid computing resources to be parallelized and obtain significant gain in computing time to be used in a near future in clinical routine for some specific applications. Currently, computing resources in EGEE, allocated to biomedical applications, consist of 1785 CPUs distributed on 23 geographic sites in Europe and Taiwan. In order to enable a transparent and interactive use of GATE applications on the grid, all the functionalities to run GATE simulations on distributed resources have been developed [8].

The GENIUS web portal [9] is implemented on top of the middleware services of the EGEE infrastructure. From the user's workstation, a web browser is running which enables the access to the EGEE grid infrastructure by having access to the User Interface (UI) machine. Using the services on the web portal, the user can interact with files on the UI and from there the user can launch jobs to the grid and manage his data. A multi-layered

security infrastructure is guaranteed to the user (transactions under SSL, user interface login and password, grid certificates). All the functionalities necessary to run GATE on the EGEE infrastructure have been encoded. The files necessary to run GATE on the grid are automatically created: the script describing the environment of computation, the GATE macros describing the simulations, the status files of the RNG (a sequence of random numbers generated by the RNG is partitioned into suitable independent subsequences) and the job description files. In order to show the advantage for the GATE simulations to partition the calculation on multiple processors, a simulation ran locally on a single processor Intel Xeon 3.06 GHz and was executed in parallel on two grid sites: CCIN2P3 (Centre de Calcul of the Institut National de Physique Nucleaire et de Physique des Particules) and LPC (Laboratoire de Physique Corpusculaire Clermont-Ferrand) . Splitting was done in 1, 10, 20, 50 and 100 subsimulations. The efficiency enhancement is defined here as a straightforward speed-up similar to AE (eq. 2) for an otherwise unloaded cluster.

3. Results

3.1. *Variance reduction techniques*

An efficiency test was performed by acquiring a ^{67}Ga-spectrum via the pulse height analyzer of a typical SPECT simulation study. Figure 1 clearly shows a manifest reduction in variance. Table 1 shows the corresponding efficiency enhancement.

Table 1
VRT comparison

	imp. sampling	analog
Activity	100MBq	100MBq
Acq. time	30s	30s
Sim. time	5,600,000s	1,954,000s
Detections	5,165,891	325,920
QF	0.92	1

The efficiency of each simulation is calculated as described in section 2.1 and in equation 1. Division of the efficiencies gives an indication of the relative efficiency enhancement, being 5.1 for the VRT case. When applying VRTs one should take into account that the statistics of the results are no longer poissonian. Several noise restoration processes are described in literature, such as combining VRT with an intermediate Bernouilli experiment [10].

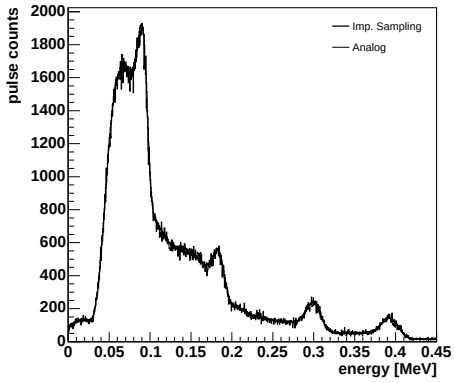

Fig. 1. ^{67}Ga spectra for simulations with and without importance sampling.

3.2. *Running GATE on a cluster*

The results of running GATE on a cluster are described in figure 2. For SPECT only a small deviation from the Amdahl prediction for linear scalability is observed, so the duration of a SPECT simulation basically drops with the number of CPUs if run on a cluster. For PET we see a large deviation from Amdahl's linear prediction. This can be explained as follows:
the SPECT benchmark is a typical example of a low sensitivity system characterized by relatively small output file sizes for a long simulation time, whereas the PET setup resembles a high sensitivity system with relatively high output volumes for a short simulation time. The percentual contributions of the output merger become larger for an increasing number of CPUs and this is a bottleneck for simulations with high data output sizes as shown in figure 3.

In a worst case scenario, high output, short duration simulation studies, can reach an optimum in cluster operation mode. That optimum can eas-

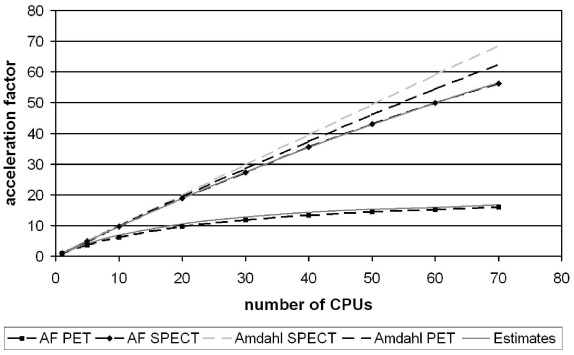

Fig. 2. Efficiency enhancements for the SPECT and PET experiments with the upper limits predicted by Amdahl's law and the estimates including the merging time.

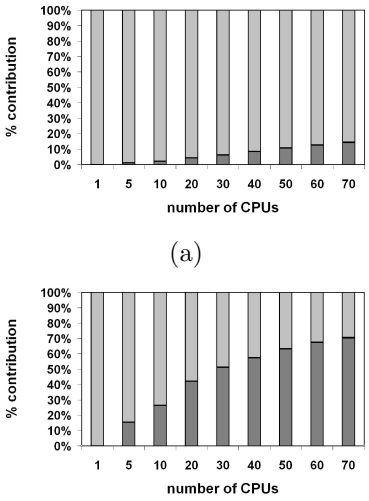

Fig. 3. Percentage of the total application time for the job splitter (black), the output merger (dark grey) and the actual simulation (light grey) as function of number of CPUs in (a) SPECT and (b) PET.

ily be calculated using equation 2 for the estimate (AE) incorporating the merging time for which we see a good agrement on figure 2.

3.3. *Deployment on the Grid*

Table 2 illustrates the computing time in minutes of a GATE simulation running on a single processor Intel Xeon 3.06 GHz locally and the same simulation split in 10,20, 50 and 100 jobs on the CCIN2P3 site. In this case the lowest computing time is obtained for 20 jobs running in parallel.

Table 2
Sequential versus grid computation time at CCIN2P3.

Number of jobs	1 (local)	10	20	50	100
Computation time (in min)	276.5	60.8	40.9	53.7	51.8
Efficiency		4.5	6.8	5.1	5.3

Concerning the same type of submission on the LPC site, Table 3 shows that the lowest computing time is obtained for the maximum splitting (100 partitions).

Table 3
Sequential versus grid computation time at LPC site.

Number of jobs	1 (local)	10	20	50	100
Computation time (in min)	276.5	37.5	22.6	13.9	9.8
Efficiency		7.4	12.2	19.9	28.3

A gain factor of 28.2 is reached for this last submission that enables the user to run his simulation in 10 minutes instead of more than 4 hours on a single processor. Performance loss and non-scalability is due to the queuing time that is really dependent on the policy of the site accepting the jobs. A possible bottleneck also remains in the network connection with the remote site for analysis of the data. Therefore, there is a trade-off between job splitting and intrinsic Grid time costs.

4. Discussion

Classical ways to decrease Monte Carlo simulation time are the application of energy and path length cuts, as well as limiting the emission angle. Here, we presented the use of geometrical importance sampling for SPECT with the same objective. Using a track history tree which includes adjustments for splitting and Russian roulette in Gate, it is possible to calculate pulse height tallies. A relative figure of merit was calculated, showing an efficiency enhancement of 5.1 by using importance sampling. In the future the optimal spliting map will be derived by one extensive forward simulation or by an inverse simulation. A maximal factor of 15 is expected. Recent work is ongo-

ing for the incorporation of forced detection into GATE and for the use of parametrized voxels. On the long term, incorporation of the fictituous cross section method [11] and of precalculated detector response [12] will be studied.

Furthermore, we have designed and developed a distributed computing framework for running GATE simulations on a cluster of computers. The partitioning scheme is based on the time-domain decomposition. The implementation is fully automatic and platform independent because the software generates a set of fully resolved macros together with an on-the-fly generated cluster submit file. The experiments have shown that GATE simulations scale very well on a cluster of homogeneous computers for the case of low sensitivity (SPECT) setups and that an optimum can be derived for high sensitivity (PET) experiments. Investigation is ongoing on how to parallellize the merger bottleneck. In [13] we already report on the use of this cluster software for the first time in research.

Finally, the parallelization of GATE simulations on the grid has shown its efficacy but depends on the queuing policy of the site accepting the jobs. The results obtained are very encouraging and with the CPU resources increasing in future Grid projects, we can expect an increased efficiency enhancement factor.

5. Conclusion

GATE is still fundamentally slower compared to analytical simulators or dedicated packages but recent efforts in efficiency enhancement have narrowed the gap significantly and will do so even more in the future. Several cluster testing sites are running realistic GATE simulations already overnight, even for observer studies.

References

[1] S. Jan, G. Santin, D. Strul, S. Staelens et al., GATE: a simulation toolkit for PET and SPECT, Physics in Medicine and Biology 49 (19) (2004) 4543–4561.

[2] M. Ljungberg, S. Strand, A Monte Carlo program for the simulation of scintillation camera characteristics, Computer Methods and Programs in Biomedicine 29 (1989) 257–272.

[3] J. De Beenhouwer, S. Staelens, M. Dressel, Y. D'Asseler et al., Geometrical importance sampling and pulse height tallies in gate, in: Proceedings of the 26th annual international conference of the IEEE EMBS, San Francisco, 2004, pp. 1349–1352.

[4] D. Haynor, R. Harrison, T. Lewellen, The use of importance sampling techniques to improve the efficiency of photon tracking in emission tomography simulations, Medical Physics 18 (5) (1991) 990–1001.

[5] J. De Beenhouwer, D. Kruecker, S. Staelens, L. Ferrer et al., Distributed computing platform for PET and SPECT simulations with GATE, accepted for the the IEEE Medical Imaging Conference (may 2005).

[6] G. Amdahl, Validity of single-processor approach to achieving large-scale computing capability, in: Proceedings of the AFIPS Conference, Reston, VA., 1967, pp. 483–485.

[7] EGEE, `http://www.eu-egee.org/`.

[8] L. Maigne, D. Hill, P. Calvat, V. Breton et al., Parallelization of monte carlo simulations and submission to a grid environment, Parallel Processing Letters Journal 14 (2) (2004) 177–196.

[9] G. Andronico, R. Barbera, A. Falzone, G. Lore et al., GENIUS: a web portal for the grid, Nuclear Instruments and Methods A 502 (2003) 433.

[10] A. Goedicke, B. Schweizer, S. Staelens, J. De Beenhouwer, Fast simulation of realistic SPECT projections using forced detection in Geant4, accepted for the EMBEC Conference (October 2005).

[11] I. Kawrakow, M. Fippel, Investigation of variance reduction techniques for Monte Carlo photon dose calculation using XVMC, Physics in Medicine and Biology 45 (8) (2000) 2163–2183.

[12] X. Song, W. Segars, Y. Du, B. Tsui et al., Fast modelling of the collimator-detector response in Monte Carlo simulation of spect imaging using the angular response function, Physics in Medicine and Biology 50 (8) (2005) 1791–1804.

[13] S. Staelens, , K. Vunckx, D. Beque, Y. D'Asseler et al., GATE multipinhole simulations : optimization of design parameters, submitted to ITBS special edition of Nuclear Instruments and Methods A (2005).